Original Paper

Improving the Prediction of Persistent High Health Care Utilizers: Retrospective Analysis Using Ensemble Methodology

Stephanie N Howson[1], MSc; Michael J McShea[1], MSc; Raghav Ramachandran[1], PhD; Howard S Burkom[1], PhD; Hsien-Yen Chang[2], PhD; Jonathan P Weiner[2], DrPH; Hadi Kharrazi[2], MD, PhD

[1]Applied Physics Laboratory, Johns Hopkins University, Baltimore, MD, United States

[2]Center for Population Health Information Technology, Johns Hopkins School of Public Health, Baltimore, MD, United States

Corresponding Author:
Hadi Kharrazi, MD, PhD
Center for Population Health Information Technology
Johns Hopkins School of Public Health
624 N Broadway
Office 606
Baltimore, MD, 21205-1900
United States
Phone: 1 443 287 8264
Email: kharrazi@jhu.edu

Abstract

Background: A small proportion of high-need patients persistently use the bulk of health care services and incur disproportionate costs. Population health management (PHM) programs often refer to these patients as persistent high utilizers (PHUs). Accurate PHU prediction enables PHM programs to better align scarce health care resources with high-need PHUs while generally improving outcomes. While prior research in PHU prediction has shown promise, traditional regression methods used in these studies have yielded limited accuracy.

Objective: We are seeking to improve PHU predictions with an ensemble approach in a retrospective observational study design using insurance claim records.

Methods: We defined a PHU as a patient with health care costs in the top 20% of all patients for 4 consecutive 6-month periods. We used 2013 claims data to predict PHU status in next 24 months. Our study population included 165,595 patients in the Johns Hopkins Health Care plan, with 8359 (5.1%) patients identified as PHUs in 2014 and 2015. We assessed the performance of several standalone machine learning methods and then an ensemble approach combining multiple models.

Results: The candidate ensemble with complement naïve Bayes and random forest layers produced increased sensitivity and positive predictive value (PPV; 49.0% and 50.3%, respectively) compared to logistic regression (46.8% and 46.1%, respectively).

Conclusions: Our results suggest that ensemble machine learning can improve prediction of care management needs. Improved PPV implies reduced incorrect referral of low-risk patients. With the improved sensitivity/PPV balance of this approach, resources may be directed more efficiently to patients needing them most.

(JMIR Med Inform 2022;10(3):e33212) doi: 10.2196/33212

KEYWORDS

persistent high utilizers; ensemble methodology; utilization; prediction; machine learning; population health analytics; retrospective; observational

Introduction

Population health management (PHM) programs regularly classify patients by estimated risk of high health care utilization such as hospitalization [1]. The classification process enables PHM programs to allocate their limited resources according to the patients' anticipated needs [1,2]. Higher-risk patient groups, if identified correctly, can receive effective interventions such as care management program enrollment to reduce utilization and improve outcomes [2]. Additionally, when utilization and costs are successfully contained for high-need patients by proactively preventing undesired outcomes, PHM programs can better allocate the remaining resources to improve the outcomes of other patients [3].

The set of high-risk patients frequently changes over time, with most patients being high-risk for a short term [4,5]. However, some high-risk patients use health care resources persistently for an extended period (eg, more than 24 months) [4-6]. These persistent high utilizer (PHU) patients generally constitute a small segment of the overall patient population but use a considerable proportion of resources in long term [4-6]. Despite the variety of approaches taken to characterize PHUs, such as adjusting for type of utilization, total costs, number of chronic conditions, and other factors, predicting who becomes a PHU has remained an analytical challenge [7-11].

Past studies have applied several analytical approaches to identify and predict PHUs in different patient populations. These approaches range from traditional regression methods (eg, logistic regression) [4-8] to complex machine learning techniques (eg, gradient boosting and neural networks) [9-11]. Nonetheless, due to the small number of PHUs in a patient population (often less than 5%), most studies have suffered from either oversensitive models or excessive false predictions of high utilization [3,5]. Thus, the challenge of achieving simultaneously useful levels of sensitivity and positive predictive value (PPV) in PHU prediction models has limited their application in practice [12].

To address the methodological challenges in predicting PHUs, this study tests an ensemble approach to balance the sensitivity and PPV of PHU forecasting at practical levels. The ensemble approach uses a mix of machine learning methodologies to improve both the sensitivity and PPV of PHU predictions at the same time. Using insurance claims data of a large patient population, this study compares the ensemble approach to single models, a baseline model, and a more advanced predictive model.

Methods

Overall Aims and Definitions

The overall goal of our study was to assess the value of ensemble methodology for achieving required levels of sensitivity and PPV for PHU prediction. Our analysis aimed to provide a methodology to optimize the tradeoff of highly sensitive and highly specific predictive models of PHUs using an ensemble approach.

We defined a PHU as an individual who remained in the top 20% of highest health care costs for 4 consecutive 6-month periods (ie, total of 24 months after the base period) [4]. Health care costs were defined as the sum of costs covered by the insurer and the patient's out-of-pocket costs [4].

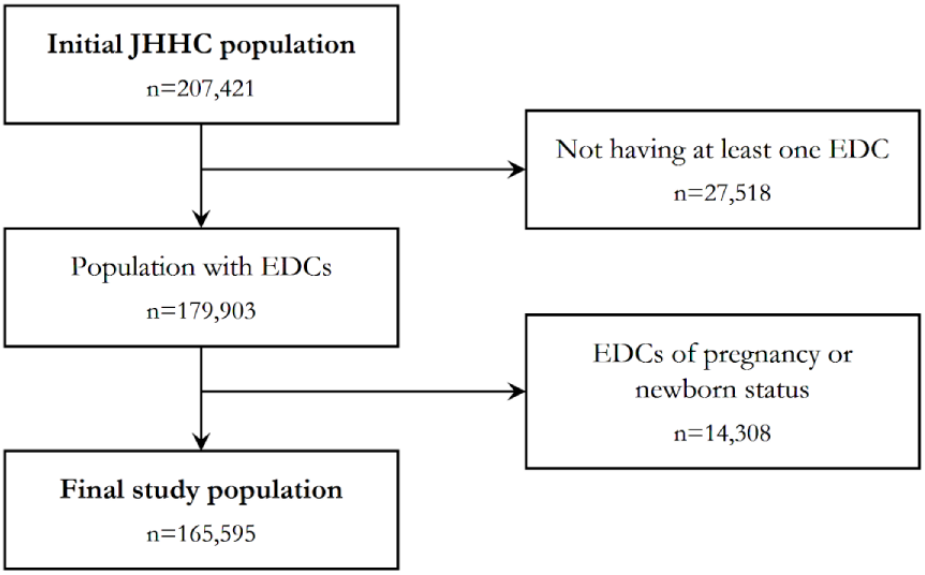
Data Source and Preparation

We performed a retrospective analysis of the Johns Hopkins Health Care insurance claims data collected between 2013 and 2015. We applied the Johns Hopkins Adjusted Clinical Groups (ACG) software to the claims data to prepare the data for analysis [13]. We categorized the diagnostic codes into higher-level diagnosis groupings called expanded diagnostic clusters (EDCs), and we grouped medication data into Rx-defined morbidity groups (RxMGs) [4,13]. EDCs and RxMGs have been substantially validated in past studies and are routinely used for risk stratification in practice [4,14].

Study Population

Johns Hopkins Health Care claims data included 207,421 patients with at least 1 record in 2013 and at least 2 years of continuous enrollment between 2013 and 2015 (Figure 1). First, 27,518 patients with missing EDC diagnosis codes were excluded, since EDCs were used to predict PHU status within the population. Second, 14,308 patients with EDC codes indicating pregnancy/newborn status were removed, as the anticipated high utilization incurred by these patients are different from PHUs. The final study population included 165,595 patients (Figure 1).

Figure 1. Selection process of the study population. JHHC: Johns Hopkins Health Care; EDC: expanded diagnostic cluster.



Predictors and Outcome

Predictors (ie, independent variables) included demographics, EDCs, RxMGs, and other health utilization variables (eg, hospitalization) generated by the ACG system. Many of these predictors, including all EDCs and RxMGs, are categorical variables [13,14].

The outcome of interest, a binary variable, was whether a patient became a PHU after the base year (ie, incurred health care costs in the top 20% of all patients over 4 consecutive 6-month periods).

Statistical Approach

Ensemble Methodology

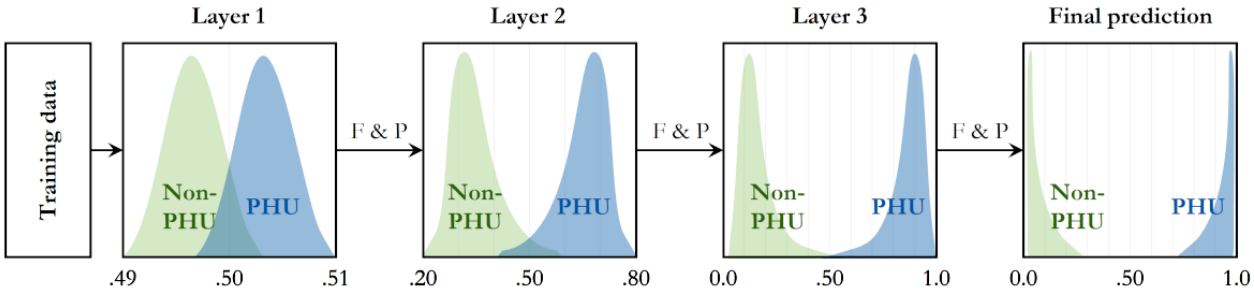
PHUs constitute a small fraction of the patient population, hence producing a large class imbalance (ie, most patients are non-PHUs). A common issue with single model prediction of highly imbalanced classes is compromising PPV in favor of higher sensitivity. For example, a single predictive model of PHUs may result in many false positives (ie, low PPV) if aiming to capture all PHUs (ie, high sensitivity). However, ensemble models provide a unique opportunity to increase both PPV and sensitivity by combining substantially different predictive models. We hypothesized that an ensemble approach can predict PHUs with both a manageable PPV and an optimal sensitivity compared to basic and advanced single model predictions.

We assessed several machine learning algorithms to predict PHU status among the study population. We also evaluated the performance of the ACG system, a comprehensive regression-based risk stratification tool commonly used in PHM practice [13]. As hypothesized, each of these algorithms yielded average levels of PPV, and we used an ensemble methodology to boost the overall PHU prediction performance.

Ensemble methods take inputs from multiple models and combine the outputs in various ways to strengthen prediction results [15]. In classification problems with imbalanced classes, ensemble methods perform well because multiple models can contribute individual strong features to the overall prediction [16]. Since PHUs make a fraction of the total population, the occurrence of a PHU in the data can be considered an anomaly [4]. Sometimes referred to as anomaly detection, the supervised machine learning problem of classifying PHUs is known as the imbalanced class problem, where the majority class (ie, non-PHUs) is much more prevalent than the minority class (ie, PHUs).

We chose the stacking ensemble model rather than the voting ensemble approach. The stacking ensemble model uses a metaclassifier to aggregate the results, but the voting ensemble model needs user-specified weights to combine the classifiers, hence adding an unpractical step [15]. Thus, for this problem space and our data set, we chose the stacking ensemble. Stacking ensemble methods often use multiple model layers and a final prediction model layer. Each layer makes predictions on the input space given. We also used an additional parameter, feature propagation. This technique allows the passing of both features and predictions through each layer of the ensemble [15]. Figure 2 depicts the overall structure of our ensemble methodology and schematically shows how multiple layers can improve PPV and sensitivity simultaneously (Figure 2).

Figure 2. Stacking ensemble architecture. F&P: feature selection and predictions; PHU: persistent high utilizer; non-PHU: nonpersistent high utilizer.



Ensemble Component Model Selection

The models selected as the layers in the ensemble method were chosen using common techniques, namely assessment of common classification algorithms and random search cross-validation for parameter tuning. Typically, machine learning models are assessed for performance and generalizability. Generalizability is difficult to quantify without large unseen data sets available for testing, but a common technique to test for overfitting is k cross-fold validation. This technique tests the machine learning model against many different subsets of data and then calculates an average of all tests. For classifying PHUs, generalizability is fundamentally important because future populations tested through these algorithms will have a large variety of differences, including demographic profiles and medical conditions. Accordingly, we employed several techniques to tune the performance and generalizability of individual models before constructing the layers of the stacking ensemble [15-17].

First, we incorporated an algorithm known as complement naïve Bayes (CNB), which often produces highly sensitive predictions when classes are imbalanced [18]. The CNB model is derived from standard multinomial naïve Bayes [18]. This model has 3 main parameters, alpha, fit prior, and norm [18]. Alpha is a Laplace smoothing parameter that adjusts the shape and fit of the multinomial distribution. This parameter shifts and forms the training distribution to characterize the multidimensional space of the data. Fit prior refines class identification when only a single class is found in the training set, which can easily occur since PHUs occur infrequently in the data set. Fitting the priors of the classifier ensures that the majority class (ie, non-PHUs)

still has some probability of not occurring, even though no other class is present in the training data. The norm parameter determines whether the training involves a second normalization of weights, an additional measure to bolster the performance on imbalanced class problems like PHU detection. Naïve Bayes models are very easy to train, so a fine-tuned parameter search was performed to find more than 1 robust CNB for use in the stacking ensemble [18].

Second, we integrated a random forest (RF) classifier in the ensemble model. An RF model is a meta-estimator that fits numerous decision tree classifiers on subsets of data features and averages results (ie, polls) to improve performance [19]. Decision trees, and by association RFs, are useful in several applications due to their explainability and ease of training. Decision trees do not require normalization and can accept categorical and numerical variables; however, a shortcoming of decision trees is their difficulty with generalization. Imprecise selection of hyperparameters will make the RF tree overly complex resulting in poor performance when facing unseen patterns [19]. Since RFs are an estimator built by decision trees, many of the parameters are carried over, although additional parameters are available for the sampling and final averaging with the RF [19].

All applicable parameters of an RF were varied through a random cross-validated grid search, but a few most notably contributed to overall performance and generalizability. These parameters include number of estimators, maximum depth, minimum samples to split, minimum samples at leaf, maximum number of features, and class weight. The number of estimators is the count of how many decision trees should be fitted to make up the RF [19]. Increasing the number of estimators typically increases generalizability but must be monitored for computational complexity. Maximum depth fixes the maximum number of levels that each tree can have, which is critical in generalizability [19]. If not set, the tree is continued until each leaf is pure, meaning the tree could learn the pattern of a single person in this population, which is not extensible to unseen populations. Minimum samples to split sets the minimum number of samples at the time of a split, ensuring that each leaf has at least n–1 samples. Minimum samples at leaf is very similar to minimum samples to split but controls samples at the leaf level. In this study, minimum samples at leaf was used to ensure edge cases (ie, unique PHU patterns) were still appropriately populated with training samples. Maximum number of features describes the method used to generate each tree which in certain use cases, taking the square root or log of the total number of features, can increase an RF's performance [19].

Class weight is the most important RF parameter for performance, although setting it can negatively impact generalizability [19]. This parameter adjusts the prior weight on the positive class, which is important for imbalanced classes, and it pushes the decision tree fits to focus more closely on the minority class, making it more robust to edge cases. Since this model was designed to detect PHUs, favoring minority instead of majority class performance was key. Using specific class weights forced the decision trees to allow for a degradation in classifying non-PHUs in favor of an increase in PHU classification. Two RF models were selected from a random search cross-validation of parameters for use in the stacking ensemble. The final stacking ensemble model integrated the CNB and RF models into one predictive model.

The final ensemble model used an 80/20 split for training and testing of the data. We performed a 5-fold cross-validation on hyperparameter search and recursive feature elimination.

Performance Metrics

Typically, positive and negative class performance are assessed equally using a metric such as F1 score. In this study, as the PHU versus non-PHU classes are unequal and the positive class would constitute an infrequent occurrence, only the positive class metrics were considered key for performance improvement. Therefore, we measured PPV and sensitivity metrics to assess performance of all models (ie, individual models and ensemble model). Both performance metrics describe the classification results for the positive class (ie, PHUs). PPV is the proportion of positive classifications that are truly PHUs. Sensitivity is the proportion of PHUs who were classified as positive.

An important consideration in any machine learning algorithm evaluation is the balance among metrics. A simple way to find an appropriate balance is to change the threshold for classification. Choosing the appropriate threshold can be difficult for health care scenarios due to the risk of incorrect classification for an individual who needs treatment (ie, false negatives). Conversely, classifying too many healthy individuals at risk could overwhelm the resources available for interventions (ie, false positives). To address this issue, we calculated and then plotted sensitivity and PPV for 50 trials at thresholds spaced evenly .05 apart. We then calculated the discrimination threshold for the ensemble model to choose the optimal threshold of the PPV versus sensitivity metrics.

Finally, we compared the PPV and sensitivity of select individual models, which achieved at least 40% performance in both metrics, with the ensemble methodology. The individual models included a logistic regression, the Johns Hopkins ACG model (out-of-box and with no further training) [13], and a standalone RF model. The ensemble model included a stacking ensemble with multiple layers combining CNB and RF models.

All analyses, including descriptive analysis, individual modeling, and ensemble approach, were performed in R (version 3.5.1, R Foundation for Statistical Computing). We used Python pandas and scikit-learn for all modeling pipeline efforts (eg, data cleaning, filtering, hyperparameter search, feature selection, and RF model). We used Python ML Ensemble for the ensemble model [20]. We used Python Yellowbrick library to visualize the classification threshold of sensitivity versus positive predictive values. We used the Johns Hopkins ACG system to produce the ACG output and measure the ACG model's performance [13].

Results

Descriptive Analyses

The study population comprised 165,595 unique patients including 8359 (5.1%) PHUs (Table 1). The PHU population's average age was more than twice that of the non-PHU population (38.51 years vs 18.79 years). PHUs included fewer males (2735/8359, 32.7%) than non-PHUs (69,683/155,862, 44.7%). As expected, PHUs had more utilization than non-PHUs (1567/8359, 18.7% vs 3891/155,862, 2.5% for inpatient visits and 8332/8359, 99.7% vs 152,199/155,862, 97.3% for outpatient visits, respectively).

Table 1. Specification of the study populations (n=165,595).

	Overall study population (n=165,595)	Non-PHU[a] population (n=155,862)	PHU population (n=8359)
Age (years), mean (SD)	19.85 (17.45)	18.79 (16.82)	38.51 (18.01)
0-17, n (%)	101,264 (61.2)	99,352 (63.7)	1459 (17.5)
18-64, n (%)	63,260 (38.2)	55,666 (35.7)	6730 (80.5)
65+, n (%)	1037 (0.6)	844 (0.5)	170 (2.0)
Sex (male), n (%)	72,974 (44.1)	69,683 (44.7)	2735 (32.7)
Race, n (%)			
White	41,492 (25.1)	38,762 (24.9)	2457 (29.4)
Black	54,207 (32.7)	50,993 (32.7)	2879 (34.4)
Other[b]	149 (0.1)	143 (0.1)	6 (<0.1)
Missing[c]	69,747 (42.1)	65,964 (42.3)	3017 (36.1)
Inpatient visits, n (%)			
0	160,035 (96.6)	151,971 (97.5)	6792 (81.3)
1-5	5430 (3.3)	3866 (2.5)	1500 (17.9)
6-10	77 (<0.1)	20 (<0.1)	54 (0.6)
11+	19 (<0.1)	5 (<0.1)	13 (0.2)
Outpatient visits, n (%)			
0	3720 (2.2)	3663 (2.4)	27 (0.3)
1-5	96,122 (58.0)	94,138 (60.4)	1234 (14.8)
6-10	33,996 (20.5)	32,317 (20.7)	1428 (17.1)
11+	31,723 (19.2)	25,744 (16.5)	5670 (67.8)

[a]PHU: persistent high utilizer.

[b]Members of known race/ethnicity not equal to Asian, Hispanic, White, or Black.

[c]Members with empty values for race.

Ensemble Model

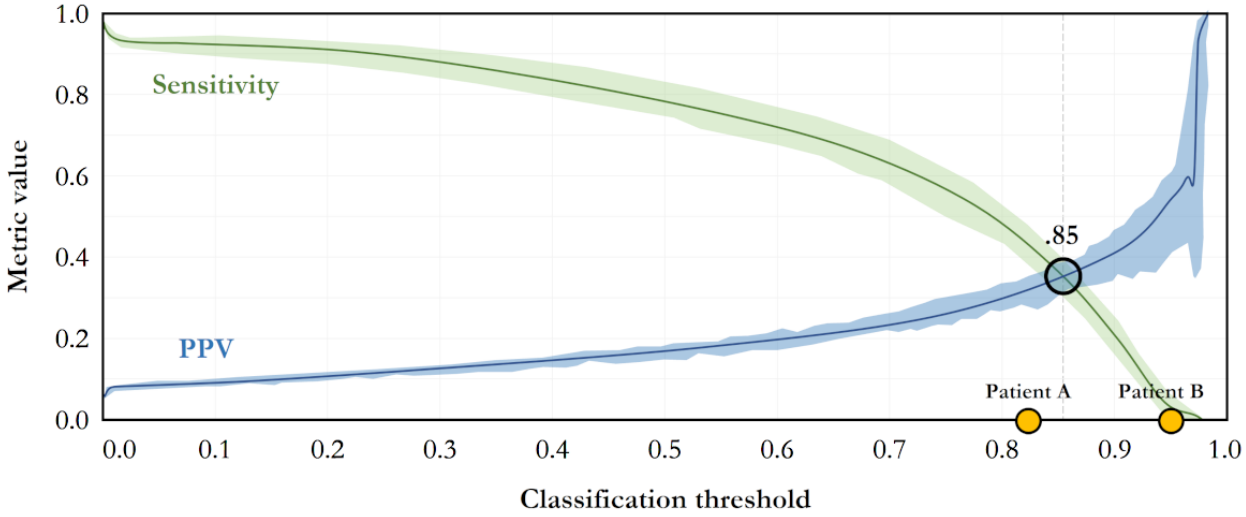
After tuning the ensemble layers, the best-performing ensemble model included 3 input layers and 1 prediction layer. The final ensemble model included 2 input layers of CNB and 1 layer of an RF model. The prediction layer was an RF model. The model included the following variables: race (ie, Black, White, other), age (as of 2013), sex, days of inpatient hospitalization in 2013, emergency department visit count in 2013, psychotherapy services in 2013, outpatient visit count in 2013, all-cause inpatient hospitalization count in 2013, frailty flag for older adults, 87 most frequent Johns Hopkins ACG diagnostic comorbidities (ie, EDCs [13]), all Johns Hopkins ACG medication grouping (ie, RxMGs [13]), and ACG-derived care coordination risk scores [13] (ie, likely coordination issue, possible coordination issue, unlikely coordination issue). These variables are generated by and included in the John Hopkins ACG risk stratification models, which are widely used for PHM efforts [13]. The stacking ensemble had full feature propagation throughout the layers to allow each model access to all data attributes while gaining classification scores from previous layers. The most performant models were selected for use in the stacking ensemble.

Model Performance Evaluation

Figure 3 depicts the discrimination threshold plot for a sample decision tree of the ensemble model. The plot conveys the importance of the threshold choice and depicts the tradeoff between PPV and sensitivity. As shown in the figure, patients A and B, both of whom are PHUs, will be identified differently by the model depending on the chosen threshold between PPV and sensitivity. By testing the trained model on these 2 patients, a risk score is generated for each. These risk scores can be compared to any classification threshold. Depending on which

side of the threshold the risk scores lie, the model classified whether patient A, B, or both are PHUs or non-PHUs.

Figure 3. Classification threshold of sensitivity versus positive predictive value (PPV): patient A: incorrectly classified as normal (risk score=82%) and patient B: correctly classified as a persistent high utilizer (risk score=97%).



The central line in Figure 3 represents the median value for each metric, and the bands represent the variability from the 10th to 90th percentiles. Two important observations about the threshold plot are (1) the typical classification threshold of .50 is not ideal probably due to the imbalanced classes and (2) equally weighting sensitivity and PPV at a threshold of .85 may not be appropriate to classify enough PHUs correctly. Patients A and B in Figure 3 have different classification outcomes and therefore interventions due in part to an arbitrary threshold.

To replicate the same level of optimality across all models, we used the 95th percentile threshold limit for each model. The absolute cutoff points were slightly different across models with ensemble having an absolute cutoff threshold of .258, RF .224, logistic regression .230, and the ACG model a cutoff of .226. Negative predictive value (NPV) and specificity were also assessed, but performance in these metrics was high (ie, averaging 97% and 99% for NPV and specificity, respectively) and did not vary significantly between models due to the large size and variability of the negative class (ie, non-PHUs).

Performance Comparison

The stacking ensemble method achieved a sensitivity of 49.0% and PPV of 50.3%. The ensemble model resulted in a 5%+ increase in both PPV and sensitivity for predicting PHUs over other individual methods such as logistic regression, RF model, and the ACG model (Table 2). As shown in Table 2, the individual RF was the highest performing nonensemble technique. Table 2 also includes the optimal parameters used in the stacking ensemble (eg, CNB and RF parameters such as alpha, maximum depth, and minimum sample splits). The final ensemble model also produced an NPV of 97.4%, specificity of 97.3%, and F1 of 49.1% for PHUs and 97.4% for non-PHUs (not shown in Table 2). The area under the curve of the ensemble model reached .921; however, comparison of areas under the curve between models was considered not valuable due to the large imbalance of PHUs versus non-PHUs, hence limiting the performance measure comparison to PPV and sensitivity of the models.

Table 2. Model fit statistics for predicting persistent high utilizer status.

Model	Parameter tuning	Sensitivity, %	PPV[a], %
Stacking ensemble Layer 1: CNB[b] Layer 2: CNB Layer 3: RF[c] Prediction layer: RF Feature propagation	CNB1 =.70, fit prior, norm CNB2 =.15, fit prior RF1 • 200 estimators • 400 max[d] depth • 5 min[e] samples split • 0.01% min samples Leaf • auto max features • class weight=0.842 RF2 • 100 estimators • 350 max depth • 2 min samples split • 0.01% min samples Leaf • class weight=1.0	49.0	50.3
RF	• 300 estimators • 500 max depth • 20 min samples split • 0.01% min samples leaf	48.4	47.2
JHU-ACG[f]	ACG[g] system probability of PHU[h]	44.7	44.1
Logistic regression	Based on 241 parameters (ie, diagnoses and medications)	46.8	46.1

[a]PPV: positive predictive value.

[b]CNB: complement naïve Bayes.

[c]RF: random forest.

[d]max: maximum.

[e]min: minimum.

[f]JHU-ACG: ACG predictive model with no local tuning.

[g]ACG: adjusted clinical group.

[h]PHU: persistent high utilizer.

Discussion

Principal Findings

Persistent high utilizers (PHUs) are defined as patients who consistently stay in the highest deciles of health care costs or utilization across multiple years [4-12]. Risk stratification efforts strive to better identify and manage PHUs so that scarce health care resources can be better allocated. Nonetheless, predicting who becomes a PHU is often challenging, partly because PHUs are uncommon [4,6,9-11]. Past studies have attempted to improve the prediction of PHUs in various populations; however, those predictions have either suffered from high false negative/positive rates or have been limited in scope [4,6,9-11]. In this study, to address the methodological complexity in predicting PHUs, we evaluated the benefit of an ensemble approach to balance the sensitivity and specificity of predicting PHUs.

Our results show that ensemble methodology can be effectively used to improve both sensitivity and PPV of predicting PHUs. The ensemble model developed in this study included 2 layers of CNB and 1 prediction layer of RF, which can be converged rather quickly. We achieved a sensitivity and PPV of 49.0% and 50.3%, respectively, using the ensemble model. In comparison to the best alternative performing model, which was the standalone RF, the ensemble model improved the sensitivity by 0.6 and PPV by 3.1 absolute percentage points, which represents a 1.2% and 6.6% relative improvement in sensitivity and PPV, respectively. Moreover, standalone RF models are prone to overfitting and often lack generalizability to other populations. The ensemble model was also superior compared to traditional logistic regression and the more established (ACG) models [13]. The ensemble model improved the sensitivity and PPV of predicting PHUs by 2.2 and 4.2 absolute percentage points (ie, 4.7% and 9.1% relative improvement) compared to the traditional logistic regression and by 4.3 and 6.2 absolute percentage points (ie, 9.6% and 14.1% relative improvement) when compared to the ACG model [13].

Several studies have examined the use of traditional methods in predicting PHUs; however, models developed in these studies have often generated low PPV rates or showed limited generalizability. For example, in a study of an employer-based health plan, using commercial claims data, a logistic regression model achieved a sensitivity of 80% but PPV of 19% to predict PHUs among the health plan enrollees [6]. In another study aiming to predict PHUs, using diagnostic and medication information extracted from claims data, a regression model achieved a sensitivity of 46.7% and PPV of 57.2%; however, the study population was limited to patients aged 18 to 62 years, hence limiting generalizability to other populations [4]. Several studies have used regression models to control for underlying demographic and clinical variables and measure the residual differences such as cost, behavioral health, and social determinants of health variables between PHU and non-PHU populations [7,8]. These studies, however, have not published the performance of these regression models in predicting PHUs.

A few studies have assessed the value of machine learning methods in predicting PHUs. In a study of a statewide Medicaid population, demographics, diagnostics, and medication information were used to predict costs associated with PHUs. The study compared multiple models including linear regression, regularized regression, gradient boosting machine, and recurrent neural networks, but the study did not generate comparable predictive measures as these models did not predict PHU status [9]. Another study applied penalized regression, support vector machine, and extreme gradient boosting against claims data to predict PHUs among patients from an academic medical center. The study achieved high sensitivity rates ranging from 72.7% to 78.7%; however, the (recalculated) PPV ranged from 18.6% to 19.8% [10]. Among the machine learning studies targeting PHUs, only one study compared an ensemble methodology (using RFs) to other methods (eg, linear regression, decision tree regression) [11]. This study, however, predicted cost of PHUs and was limited to patients with schizophrenia, hence limiting its generalizability to the broader population of patients.

Despite the promising findings of past studies in predicting PHUs, their results cannot be accurately compared to our ensemble model as each study used a slightly different definition of PHU. Some studies have defined PHUs as patients in the top 5% of cost over 2 years [4], while other studies have set the bar at 10% or 20% of cost over longer periods of time [6,7]. Future research should attempt to harmonize the definition of PHUs to make the comparison of PHU populations across different populations and health plans feasible. Additionally, harmonization of the PHU definition can facilitate the performance measurement and comparison of PHU predictive models across different health care settings.

Balancing the sensitivity and PPV of PHU predictions is key in operationalizing such models in PHM efforts. Indeed, given the infrequency of PHUs in the total population of patients, a balanced sensitivity and PPV ratio will play an important role in the management of limited resources for PHUs. In our study, the improvement of model performance compared to the traditional models corresponds to approximately 84 additional PHUs being classified correctly in the test set of 1672 true PHUs. These 84 patients would not have been reviewed for potential proactive interventions by a care manager if tested by a traditional method.

In this study, we chose to report classification performance at the balanced precision and recall scores (50/50) to highlight optimal performance in both metrics simultaneously. In specific PHM use cases, it may be desirable to select a lower classification threshold and more patients for care or intervention consideration, even if their individual risk score is lower. In large-scale PHM use cases, cost of considering many patients may be too high and a higher classification threshold is to be selected to only manage the most at-risk patients. Hence, individual population health programs may chose different balances of precision versus recall for models predicting PHUs.

Our study showed that machine learning has a performance advantage over traditional statistical models. Ultimately, improved performance will come from more advanced ensemble methods coupled with continually improving robustness of feature analysis, which together are the keys to significantly increased performance. Model performance could benefit from subpopulation training by reducing the large and variable parameter space for classification. Thus, developing custom groupings of clinical features associate with PHU patients (versus non-PHUs) can potentially advance predictive models of PHUs. For example, clinical groupings identified by unsupervised machine learning techniques (such as latent class analysis) has shown value in improving predictive models of PHUs [21].

Value-based health care providers are increasingly using risk stratification tools to manage their patient populations [22]. Providers often use local electronic health records (EHRs) instead of insurance claims to risk stratify patients and predict PHUs [23-25]. Although advances have been made in using unique EHR data to improve risk prediction using prescription data [26-28], vital signs [29,30], laboratory results [31], and free-text analysis [32,33], quality of EHR data remains a major challenge in this process [34]. Using machine learning models, such as the ensemble models, can potentially help providers address some of these deficiencies and improve the prediction of PHUs using EHR data [35,36]. Future studies should investigate the usability of machine learning models in enhancing EHR-based PHU predictions and its implication on improving the wider population-level health outcomes [37].

Limitations

Our study has several limitations. First, the results of our ensemble approach and the improvement of the PHU prediction may not generalize to other populations (eg, older adults), different settings (eg, inpatient only), or alternative data sources (eg, EHRs). Future research should explore the use of ensemble methodology in new populations and settings using alternate data sources. Second, the current definition of PHU may not be consistent with the operational definition in all PHM. We used a specific definition for PHU (ie, percentile of cost and time period), but that definition may not fit all populations. The risk stratification research community should harmonize the definition of PHU so predictive models of PHUs can be compared accurately to increase their generalizability. Third, we only used demographics, diagnosis, and medications in our

prediction models. Past research has shown the value of social determinants of health in improving the prediction of health care utilization [38-42]. Future research should investigate the value of the ensemble model in improving predictive models of PHU that incorporate social data. Finally, the ensemble methodology uses an approach that complicates the explanation of a prediction, and thus the operational use of such models in clinical and PHM settings should be further studied.

Conclusion

A small segment of the patient population uses most of the health care services over extended periods. We used an ensemble model, a machine learning approach that combines multiple modeling techniques, to simultaneously improve the sensitivity and PPV of predicting PHUs using claims data. Future studies should investigate the value of machine learning techniques in predicting PHUs in other health care settings with potentially different underlying populations and different data sources (eg, EHR data).

Acknowledgments

We acknowledge the contributions of Sheri Maxim, Jonathan Thornhill, Jason Lee, Hong Kan, and Tom Richards to this project. This project was funded by the Johns Hopkins Applied Physics Laboratory's National Health Mission Area Independent Research and Development program.

Authors' Contributions

HK and MJM codirected the research project. SNH analyzed the data. HYC provided analytical insight and calculated claims costs. HK, MJM, HSB, RR, and JPW reviewed and interpreted the results. HK, SNH, and MJM drafted the manuscript. All authors reviewed and contributed to the final manuscript. HK prepared the manuscript for submission.

Conflicts of Interest

None declared.

References

1. Iezzoni LI. Risk Adjustment for Measuring Health Care Outcomes, Fourth Edition. New York: Health Administration Press; 2012.
2. Kharrazi H, Gamache R, Weiner J. Role of informatics in bridging public and population health. In: Magnuson J, Dixon B, editors. Public Health Informatics and Information Systems. London: Springer; 2020.
3. Lee NS, Whitman N, Vakharia N, Ph DBT, Rothberg MB. High-cost patients: hot-spotters don't explain the half of it. J Gen Intern Med 2017 Jan;32(1):28-34 [FREE Full text] [doi: 10.1007/s11606-016-3790-3] [Medline: 27480529]
4. Chang H, Boyd CM, Leff B, Lemke KW, Bodycombe DP, Weiner JP. Identifying consistent high-cost users in a health plan: comparison of alternative prediction models. Med Care 2016 Sep;54(9):852-859. [doi: 10.1097/MLR.0000000000000566] [Medline: 27326548]
5. Guilcher SJT, Bronskill SE, Guan J, Wodchis WP. Who are the high-cost users? A method for person-centred attribution of health care spending. PLoS One 2016;11(3):e0149179 [FREE Full text] [doi: 10.1371/journal.pone.0149179] [Medline: 26937955]
6. Hwang W, LaClair M, Camacho F, Paz H. Persistent high utilization in a privately insured population. Am J Manag Care 2015 Apr;21(4):309-316 [FREE Full text] [Medline: 26014469]
7. Yoon J, Chee CP, Su P, Almenoff P, Zulman DM, Wagner TH. Persistence of high health care costs among VA patients. Health Serv Res 2018 Oct;53(5):3898-3916 [FREE Full text] [doi: 10.1111/1475-6773.12989] [Medline: 29862504]
8. Sterling S, Chi F, Weisner C, Grant R, Pruzansky A, Bui S, et al. Association of behavioral health factors and social determinants of health with high and persistently high healthcare costs. Prev Med Rep 2018 Sep;11:154-159 [FREE Full text] [doi: 10.1016/j.pmedr.2018.06.017] [Medline: 30003015]
9. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed Eng Online 2018 Nov 20;17(Suppl 1):131 [FREE Full text] [doi: 10.1186/s12938-018-0568-3] [Medline: 30458798]
10. Ng SHX, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore. BMJ Open 2020 Jan 06;10(1):e031622 [FREE Full text] [doi: 10.1136/bmjopen-2019-031622] [Medline: 31911514]
11. Wang Y, Iyengar V, Hu J, Kho D, Falconer E, Docherty JP, et al. Predicting future high-cost schizophrenia patients using high-dimensional administrative data. Front Psychiatry 2017;8:114 [FREE Full text] [doi: 10.3389/fpsyt.2017.00114] [Medline: 28713293]
12. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. CMAJ 2016 Feb 16;188(3):182-188 [FREE Full text] [doi: 10.1503/cmaj.150064] [Medline: 26755672]
13. Johns Hopkins ACGs System, version 12. Johns Hopkins School of Public Health. 2019. URL: https://www.hopkinsacg.org/ [accessed 2022-02-07]

14. Weiner JP, Starfield BH, Steinwachs DM, Mumford LM. Development and application of a population-oriented measure of ambulatory care case-mix. Med Care 1991 May;29(5):452-472. [doi: 10.1097/00005650-199105000-00006] [Medline: 1902278]
15. Zhi-Hua Z. Ensemble Methods: Foundations and Algorithms, 1st Edition. New York: Chapman and Hall/CRC; 2012.
16. Chen Z, Duan J, Kang L, Qiu G. Class-imbalanced deep learning via a class-balanced ensemble. IEEE Trans Neural Netw Learn Syst 2021 Apr 26:1. [doi: 10.1109/TNNLS.2021.3071122] [Medline: 33900923]
17. Yu K, Xie X. Predicting hospital readmission: a joint ensemble-learning model. IEEE J Biomed Health Inform 2020 Feb;24(2):447-456. [doi: 10.1109/JBHI.2019.2938995] [Medline: 31484143]
18. Rennie J, Shih L, Teevan J, Karger D. Tackling the poor assumptions of naive Bayes text classifiers. Proc 20th Int Conf Mach Learn 2003;3:616-623.
19. Breiman L. Random forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
20. Flennerhag S. ML-Ensemble: high performance ensemble learning in Python. URL: http://ml-ensemble.com/ [accessed 2022-02-09]
21. Ramachandran R, McShea MJ, Howson SN, Burkom HS, Chang H, Weiner JP, et al. Assessing the value of unsupervised clustering in predicting persistent high health care utilizers: retrospective analysis of insurance claims data. JMIR Med Inform 2021 Nov 25;9(11):e31442 [FREE Full text] [doi: 10.2196/31442] [Medline: 34592712]
22. Pandya CJ, Chang H, Kharrazi H. Electronic health record-based risk stratification: a potential key ingredient to achieving value-based care. Popul Health Manag 2021 Jun 14;24(6):654-656. [doi: 10.1089/pop.2021.0131] [Medline: 34129398]
23. Kharrazi H, Chi W, Chang H, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. Med Care 2017 Aug;55(8):789-796. [doi: 10.1097/MLR.0000000000000754] [Medline: 28598890]
24. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. Med Care 2018 Dec;56(2):202-203. [doi: 10.1097/MLR.0000000000000849] [Medline: 29200132]
25. Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. J Med Internet Res 2018 Aug 07;20(8):e10458 [FREE Full text] [doi: 10.2196/10458] [Medline: 30087090]
26. Chang H, Richards TM, Shermock KM, Elder Dalpoas S, J Kan H, Alexander GC, et al. Evaluating the impact of prescription fill rates on risk stratification model performance. Med Care 2017 Dec;55(12):1052-1060. [doi: 10.1097/MLR.0000000000000825] [Medline: 29036011]
27. Kharrazi H, Ma X, Chang H, Richards TM, Jung C. Comparing the predictive effects of patient medication adherence indices in electronic health record and claims-based risk stratification models. Popul Health Manag 2021 Oct;24(5):601-609. [doi: 10.1089/pop.2020.0306] [Medline: 33544044]
28. Chang H, Kan HJ, Shermock KM, Alexander GC, Weiner JP, Kharrazi H. Integrating e-prescribing and pharmacy claims data for predictive modeling: comparing costs and utilization of health plan members who fill their initial medications with those who do not. J Manag Care Spec Pharm 2020 Oct;26(10):1282-1290. [doi: 10.18553/jmcp.2020.26.10.1282] [Medline: 32996394]
29. Kharrazi H, Chang H, Heins SE, Weiner JP, Gudzune KA. Assessing the impact of body mass index information on the performance of risk adjustment models in predicting health care costs and utilization. Med Care 2018 Dec;56(12):1042-1050. [doi: 10.1097/MLR.0000000000001001] [Medline: 30339574]
30. Kharrazi H, Chang H, Weiner JP, Gudzune KA. Assessing the added value of blood pressure information derived from electronic health records in predicting health care cost and utilization. Popul Health Manag 2021 Nov 29:250. [doi: 10.1089/pop.2021.0250] [Medline: 34847729]
31. Lemke KW, Gudzune KA, Kharrazi H, Weiner JP. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. Am J Manag Care 2018 Jun 01;24(6):e190-e195 [FREE Full text] [Medline: 29939509]
32. Kan HJ, Kharrazi H, Leff B, Boyd C, Davison A, Chang H, et al. Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. Med Care 2018 Dec;56(3):233-239. [doi: 10.1097/MLR.0000000000000865] [Medline: 29438193]
33. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
34. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. J Gen Intern Med 2014 Jul;29(7):976-978 [FREE Full text] [doi: 10.1007/s11606-014-2883-0] [Medline: 24839057]
35. Ng SH, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. BMC Health Serv Res 2019 Jul 05;19(1):452 [FREE Full text] [doi: 10.1186/s12913-019-4239-2] [Medline: 31277649]

36. Kan HJ, Kharrazi H, Chang H, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One 2019;14(3):e0213258 [FREE Full text] [doi: 10.1371/journal.pone.0213258] [Medline: 30840682]
37. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. Yearb Med Inform 2018 Aug;27(1):199-206 [FREE Full text] [doi: 10.1055/s-0038-1667081] [Medline: 30157524]
38. Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. Popul Health Manag 2021 Apr;24(2):222-230. [doi: 10.1089/pop.2020.0043] [Medline: 32598228]
39. Chang H, Hatef E, Ma X, Weiner JP, Kharrazi H. Impact of area deprivation index on the performance of claims-based risk-adjustment models in predicting health care costs and utilization. Popul Health Manag 2021 Jun;24(3):403-411. [doi: 10.1089/pop.2020.0135] [Medline: 33434448]
40. Hatef E, Kharrazi H, Nelson K, Sylling P, Ma X, Lasser EC, et al. The association between neighborhood socioeconomic and housing characteristics with hospitalization: results of a national study of veterans. J Am Board Fam Med 2019;32(6):890-903 [FREE Full text] [doi: 10.3122/jabfm.2019.06.190138] [Medline: 31704758]
41. Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, et al. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. JMIR Med Inform 2020 Sep 08;8(9):e18084 [FREE Full text] [doi: 10.2196/18084] [Medline: 32897240]
42. Vest JR, Adler-Milstein J, Gottlieb LM, Bian J, Campion TR, Cohen GR, et al. Assessment of structured data elements for social risk factors. Am J Manag Care 2022 Jan 01;28(1):e14-e23 [FREE Full text] [doi: 10.37765/ajmc.2022.88816] [Medline: 35049262]

Abbreviations

ACG: Adjusted Clinical Group
CNB: complement naïve Bayes
EDC: expanded diagnostic cluster
EHR: electronic health record
NPV: negative predictive value
PHM: population health management
PHU: persistent high utilizer
PPV: positive predictive value
RF: random forest
RxMG: Rx-defined morbidity group

Edited by C Lovis; submitted 27.08.21; peer-reviewed by J Coquet, S Nagavally; comments to author 20.09.21; revised version received 21.02.22; accepted 11.03.22; published 24.03.22

Please cite as:
Howson SN, McShea MJ, Ramachandran R, Burkom HS, Chang HY, Weiner JP, Kharrazi H
Improving the Prediction of Persistent High Health Care Utilizers: Retrospective Analysis Using Ensemble Methodology
JMIR Med Inform 2022;10(3):e33212
URL: https://medinform.jmir.org/2022/3/e33212
doi: 10.2196/33212
PMID: 35275063